EJERCICIO CON ORDENADOR II.B: simulación de un modelo

a) Generar mediante un programa informático el modelo estadístico de regresión lineal simple

yi=3+2xi+N(0,1), i=1,…100,

siendo x'= (1,…1,2,…,2,…,10,…,10), repitiéndose 10 veces cada dígito del 1 al 10.

b) Estudiar el ajuste del modelo, obteniendo una estimación del intercepto y la pendiente., junto con la gráfica de la nube de puntos y la recta de regresión,

c) Aumentar el tamaño de la muestra a n=500, apareciendo 50 veces los dígitos del 1 al 10. ¿Qué cambios se observan?

SOLUCIÓN

a)

Vamos a generar primero los datos en R (no en Rcmdr)

Generamos el vector x de variable explicativas

> x<-as.numeric(gl(10,10))

gl(k,n) : generador de niveles, k niveles y n número de veces que se repiten, es un vector de **factores** (datos que son variables categóricas cualitativas)**;** por defecto, como niveles toma números consecutivos a partir del 1

as.numeric: cambia el vector de factores a **numéricos**, con los que ya se puede operar.

(OBSERVACIÓN: a R se le puede preguntar por la sintaxis de un comando,mediante >?nombre comando,

p. ej., >?gl)

Verificamos que hemos generado el vector x

```
> x
 [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3
[26] 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5
[51] 6 6 6 6 6 6 6 6 6 6 7 7 7 7 7 7 7 7 7 7 8 8 8 8 8
[76] 8 8 8 8 8 9 9 9 9 9 9 9 9 9 9 10 10 10 10 10 10 10 10 10 10
```

Generamos el vector de la variable explicada

```
> y<-3+2*x+rnorm(100,mean=0,sd=1)
```

rnorm: generación aleatoria de una muestra de datos (en nuestro caso de tamaño 100) de una población normal con esa esperanza y desviac. típica.

Generamos la tabla de datos con columnas x y al que asignamos un nombre

```
> model_1<-data.frame(x,y)
```

Verificamos que lo tenemos

```
> model_1
   x        y
1  1 3.699978
2  1 5.582542
3  1 6.407326
4  1 4.622501
5  1 5.867662
6  1 5.798996
7  1 6.440460
```

8 1 4.524640

9 1 5.251577

10 1 6.500915

11 2 6.918120

12 2 7.632662

13 2 8.301854

14 2 6.758904

15 2 7.490620

16 2 6.090097

17 2 7.301787

18 2 6.215211

19 2 7.232093

20 2 7.485491

21 3 7.944475

22 3 9.047126

23 3 9.916594

24 3 7.722007

25 3 8.302501

26 3 10.090432

27 3 9.958587

28 3 8.697078

29 3 7.146964

30 3 8.826056

31 4 10.471912

32 4 11.110121

33 4 11.276995

34 4 11.620970

35 4 11.350569

36 4 9.798395

37 4 9.621599

38 4 11.657241

39 4 11.207038

40 4 11.111237

41 5 13.092974

42 5 14.989702

43 5 12.222826

44 5 13.142857

45 5 13.735631

46 5 11.537763

47 5 13.530690

48 5 13.525069

49 5 14.092046

50 5 13.504156

51 6 16.822855

52 6 13.582771

53 6 14.410690

54 6 15.570049

55 6 15.544798

56 6 16.611782

57 6 14.351448

58 6 16.230820

59 6 17.121237

60 6 13.598955

61 7 17.846082

62 7 18.643253

63 7 17.419862

64 7 17.233715

65 7 16.757392

66 7 16.267113

67 7 18.191387

68 7 17.286575

69 7 16.482403

70 7 17.907894

71 8 18.790566

72 8 19.920827

73 8 17.446903

74 8 19.097690

75 8 20.034466

76 8 18.436377

77 8 20.311117

78 8 19.039673

79 8 19.967183

80 8 18.631330

81 9 20.051170

82 9 22.245064

83 9 21.404305

84 9 21.048093

85 9 20.927358

86 9 21.075704

87 9 20.951765

88 9 20.002641

89 9 20.376834

90 9 20.921058

91 10 23.970900

92 10 22.215346

93 10 23.227968

94 10 23.683089

95 10 21.595172

96 10 22.369021

97 10 22.716457

98 10 23.041267

99 10 23.898671

100 10 23.526403

Guardamos en nuestro ordenador la tabla de datos como texto (que lo leerá bien Rcmdr):

> write.table(model_1,file="datos_ejerc_IIB_b.txt")

(en windows lo guarda, por defecto, en el directorio documentos, si no está ahí, se busca)

Un vez tenemos el fichero, lo pasamos al directorio deseado (por ej. en escritorio) para abrir desde el Rcmdr.

b) En la consola de Rcmdr, importamos el fichero datos_ejerc_IIB_b.txt:

Datos -> importar datos -> desde archivo de texto etc, se le ha de dar un nombre al fichero de datos, que puede ser el mismo.

Se verifica que la matriz de datos es correcta con visualizar datos.

Estadísticos->ajuste de modelos-> regresión lineal

Obtenemos:

```
Call:

lm(formula = y ~ x, data = datos_ejerc_ord_II_b)


Residuals:

    Min      1Q  Median      3Q     Max

-2.03819 -0.68336  0.03441  0.60744  1.97610


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  3.22517    0.19138   16.85   <2e-16 ***

x            1.98666    0.03084   64.41   <2e-16 ***

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.8859 on 98 degrees of freedom

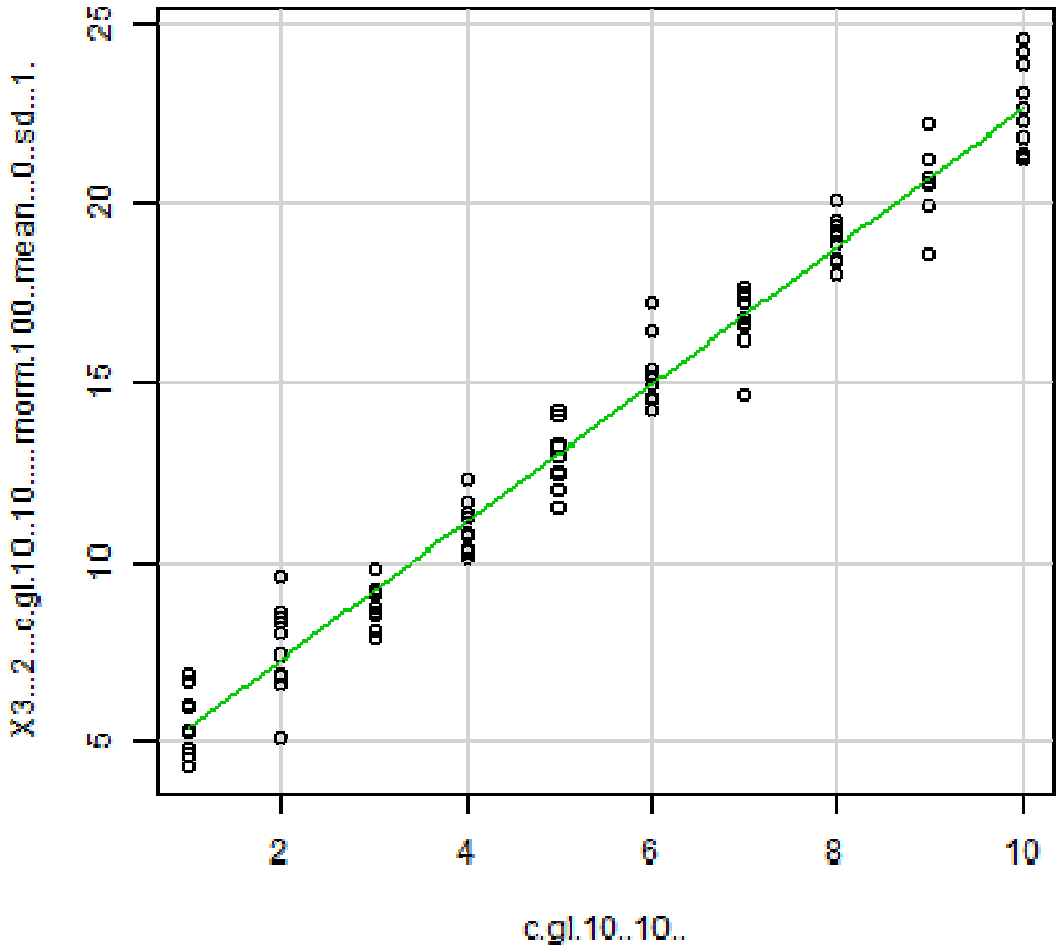
Multiple R-squared: 0.9769, Adjusted R-squared: 0.9767

F-statistic: 4149 on 1 and 98 DF, p-value: < 2.2e-16

Es fácil interpretar los resultados. Como era de esperar el p-valor sale muy pequeño, <2e-16 , de hecho en la práctica nulo, lo que nos indica que la variable y depende realmente de la variable explicativa x (la pendiente es claramente no nula). En este caso, esto tiene que ver con que el modelo se ha construido de forma que está totalmente bien especificado (¡es correcto por construcción!).

La gráfica se obtiene:

Gráficas-> diagrama de dispersión-> se eligen las variables -> opciones: se marca" línea de minimos cuadrados"



OBSERVACIÓN: SI REALIZAMOS EL EJERCICIO DE NUEVO (COMPARAR CON RESULTADOS DE OTROS ALUMNOS), NO OBTENEMOS EXACTAMENTE LOS MISMOS RESULTADOS (hay pequeñas variaciones), ¿POR QUÉ?

Dejamos que el alumno que haga el apartado c).